Weak measurement invariance with respect to unmeasured variables: An implication of strict factorial invariance.

Lubke, G.H.; Dolan, C.V.; Kelderman, H.; Mellenbergh, G.J.

published in
British Journal of Mathematical and Statistical Psychology
2003

DOI (link to publisher)
10.1348/000711003770480020

document version
Publisher's PDF, also known as Version of record

British Journal of Mathematical and Statistical Psychology (2003), 56, 231–248

www.bps.org.uk

Weak measurement invariance with respect to unmeasured variables: An implication of strict factorial invariance

Gitta H. Lubke[1]*, Conor V. Dolan[2], Henk Kelderman[3] and Gideon J. Mellenbergh[2]

[1]University of California, Los Angeles, CA, USA
[2]University of Amsterdam, The Netherlands
[3]Free University Amsterdam, The Netherlands

Measurement bias refers to systematic differences across subpopulations in the relation between observed test scores and the latent variant underlying the test scores. Comparisons of subpopulations with the same score on the latent variable can be expected to have the same observed test score. Measurement invariance is therefore one of the key issues in psychological testing. It has been established that strict factorial invariance (SFI) with respect to a selection variable *V* almost certainly implies weak measurement invariance with respect to *V*: given SFI, means and variances of observed scores do not depend on *V*. It is shown that this result can be extended. SFI in groups derived by selection on *V* has implications not only for *V* but also for potentially biasing variables *W*, if *W* and the selection variable *V* and/or if *W* and the factor underlying the observed test scores are statistically dependent. Given SFI with respect to *V* and prior knowledge concerning these dependencies, it is not necessary to measure and model variables *W* in order to exclude them as potentially biasing variables if the investigation focuses on groups selected on *V*.

1. Introduction

Establishing measurement invariance is of primary interest in psychological testing. The probability of correctly answering an item measuring, say, mathematical achievement, should be the same for all test takers with the same level of mathematical achievement, and should not depend on the test taker's other characteristics such as sex, race, or

*Requests for reprints should be addressed to Gitta H. Lubke, UCLA/GSEIS, Moore Hall, Box 951521, Los Angeles, CA, 90095-1521, USA.

parents' education. As a consequence, considerable research has focused on bias detection, that is, absence of measurement invariance (see, for example, Mellenbergh, 1989; Millsap & Everson, 1993; Shealy & Stout, 1993). Oort (1992) has emphasized the necessity of explicitly specifying the variable with respect to which a test is unbiased. If a test is unbiased with respect to, say, race, this cannot be taken to mean that the test is necessarily unbiased with respect to some other variable (Meredith, 1993). For instance, it is possible that, conditional on ability, a test item is equally difficult for children belonging to different ethnic groups, but biased with respect to sex. Ideally, then, one would have to investigate absence of bias with respect to all characteristics of the test-taking population which are deemed relevant in order to guarantee bias-free results. The present paper aims to show that such an effort can be largely reduced. To ascertain bias-free results across given groups, it is not necessary to investigate absence of bias in these groups with respect to all potentially biasing variables separately.

The paper is confined to the context of the common factor model where observed multivariate normally distributed variables are linearly regressed on underlying multivariate normally distributed factors. We consider a population consisting of several subpopulations which are defined by a grouping or selection variable V. If V is sex, then subpopulations will be boys and girls. Suppose that a test, say a mathematical achievement test, is administered to these two subpopulations and that the observed stores on the test can be adequately modelled with a common factor model. Define a potentially biasing variable W as a variable with a direct effect on (one of) the test items, say test item k.[1] This means that the regression of the test item k on W is non-zero. Bias (i.e. absence of measurement invariance) occurs if W causes mean and/or variance differences in item k between subpopulations defined by selection on V. These mean and/or variance differences between subpopulations derived by selection on V can only be due to W if the mean and/or variance of W conditional on V differs across these subpopulations. In our example, for bias to occur, (one of) the observed test items of the mathematical achievement test is influenced by a potentially biasing variable W, say attitude to mathematics, and boys and girls have to differ with respect to the mean and/or variance of attitude to mathematics and/or the strength of the influence of attitude to mathematics on the mathematical achievement item.

The test is said to be weakly measurement invariant across the subpopulations derived by selection on V (e.g. across boys and girls) if the means and (co)variances of observed test scores conditional on the selection variable and the latent variables underlying the test equal the means and (co)variances of the observed test scores conditional on the latent variables underlying the test (Meredith, 1993). In terms of the example, sex influences the means or variances of the mathematical achievement test scores only through the mean and variance of the mathematical achievement factor but sex has no direct effect on (one of) the items. Hence, sex can be excluded as a biasing variable. We argue that once weak measurement invariance has been established in groups selected on V, it is possible not only to exclude V as a biasing variable, but also to exclude other potentially biasing variables W in the groups selected on V if W and the selection variable V and/or if W and the latent variable underlying the test are statistically dependent. In our example, if weak measurement invariance of the mathematical achievement test has been established across boys and girls, then it is possible to exclude sex as a biasing variable and, in addition, attitude towards mathematics if the latter can be assumed to be correlated with sex. Note that this

[1] *For ease of presentation we consider a single item, but our argument holds for more than one item.*

extension of the theory provided by Meredith (1993) only concerns the groups selected on V (e.g. boys and girls). We do not make any statements concerning measurement invariance in groups selected on other potentially biasing variables (e.g. different maths attitude groups).

The argument is based on the concept of strict factorial invariance (SFI), which has been developed within the context of the common factor model by Meredith (1964, 1993); see also Bloxom (1972) and Ellis (1993). SFI means that the relation between test items or subscales and the underlying latent construct(s) can be represented by the same common factor model in different groups. More specifically, a test is defined to be strictly factorial invariant with respect to some variable V if the factor model, which holds in the parent population, also holds in each subpopulation derived by selection on V. The only difference between subpopulations, then, pertains to the means and (co)variances of the factors. Meredith (1993) has shown that tenability of SFI across groups derived by selection on V almost certainly implies weak measurement invariance with respect to V. We elaborate on this result and show that tenability of SFI with respect to V has implications not only for V but also for potentially biasing variables W. Consider the case where W and V are statistically dependent. Non-random selection on V introduces mean and/or variance differences in W across at least some selected subpopulations. We show that in that case SFI cannot hold in subpopulations derived by selection on V. Hence, if SFI is true, the expected value and/or the variance of W conditional on V has to be constant across subpopulations derived by selection on V. Now consider the case where W and V are statistically independent, but that W and the factor underlying the observed scores are dependent. Selection on V leaves the conditional expectance and variance of W invariant across subpopulations. We show that in this case SFI also cannot hold. Hence, if SFI is true, W and the factor cannot be dependent. Finally, consider the case where W and V, and W and the factor are independent. Again, there are no mean and variance differences in W across subpopulations selected on V. We show that SFI holds with respect to V. Mean and variance in the observed variable with the non-zero regression on W may be partially due to W. These differences are absorbed by the regression intercept and the residual variance of the regression of that variable on the factor. Since there are no subpopulation differences in W, the regression intercept and the residual variance remain invariant across subpopulations derived by selection on V, and SFI holds.

A consequence of these implications of SFI with respect to V is that if prior knowledge is present establishing the dependence of potentially biasing variables on the selection variable V and/or the factor underlying observed test scores, it is unnecessary to measure these variables and conduct bias investigations with respect to them. Given tenability of SFI with respect to V, it is possible to conclude that a variable W does not induce bias in groups selected on V without even measuring the variable, if W and the selection variable, or W and the underlying factor are known to be statistically dependent.

2. Absence of measurement bias

Various forms of bias can be distinguished. Millsap (1995, 1997, 1998) discusses the relation between predictive bias and measurement bias. Predictive bias concerns systematic group differences in the prediction of a criterion variable from test scores. Measurement bias, on the other hand, refers to group differences in the relation

between test scores and the underlying latent variable or factor. The present paper is confined to measurement bias. Absence of measurement bias with respect to a selection variable V has been defined by Mellenbergh (1989) in terms of the equality

$$f(\mathbf{Y} \mid \eta, v) = f(\mathbf{Y} \mid \eta), \tag{1}$$

where $f(\cdot)$ denotes the probability distribution function, η is a q-dimensional vector of scores on latent variables or factors underlying the p-dimensional random variable $\mathbf{Y}$, $p > q$, and v is a realization of the selection variable V.[2] This equality holds if and only if $\mathbf{Y}$ and V are conditionally (i.e. locally) independent given the factor scores η (for a proof, see Meredith, 1993, p. 528). Suppose again V represents sex, $\mathbf{Y}$ contains scores on items of a mathematical achievement test, and η is also unidimensional and represents a mathematical achievement factor. Local independence of $\mathbf{Y}$ and V given η means that the scores $\mathbf{Y}$ of male and female test takers depend solely on their level of mathematical achievement and not on their sex, that is, $\mathbf{Y}$ is not biased with respect to sex.

Equation (1) coincides with Meredith's definition of measurement invariance. Meredith's definition of weak measurement invariance (WMI) states that only the first two moments of $f(\mathbf{Y} \mid \eta, v)$ depend solely on η. Using the above example, WMI means that the means and (co)variances of test scores depend solely on mathematical achievement and not on sex. If $f(\mathbf{Y} \mid \eta, v)$ is a multinormal distribution, the distinction between measurement invariance and WMI disappears.

Analogously to Mellenbergh (1982, 1989), we distinguish between uniform and non-uniform measurement bias. Uniform bias concerns a main effect of the biasing variable: conditional on η, members of one group score consistently higher than members of another group. If the scores in both groups can be modelled in terms of linear regressions of observed scores $\mathbf{Y}$ on latent scores η, uniform bias implies group differences in the intercepts of the regression and parallel regression slopes. Non-uniform bias, on the other hand, pertains to group differences in regression slopes and can be conceptualized as an interaction of the latent variable and the biasing variable. Note that not only differences in the intercept and slope are critical in the context of measurement bias. Differences in residual variances may have important consequences even if uniform and/or non-uniform bias is absent. Suppose, for instance, that a test is used for an admission decision and a certain level of ability (e.g. the latent variable or factor) is required. If the decision is based on the observed test scores, then the number of false admissions and false rejections is higher in the group with the larger residual variance (see also Meredith, 1993, p. 530). In short, given linearity of the regression of $\mathbf{Y}$ on η, the different forms of measurement bias involve differences in intercepts, slopes and residual variances.

3. Strict factorial invariance

The usefulness of SFI as a tool in bias investigations lies in the fact that SFI implies a number of restrictions on the regression of $\mathbf{Y}$ on η. The restricted regression model, in turn, has implications with respect to measurement invariance. As we will see, the restrictions encompass equality of intercepts, slopes and residual variances. The

[2] *Without loss of generality, we assume V to be unidimensional throughout.*

regression model is the common factor model (Jöreskog, 1971),

$$y = \nu + \Lambda\eta + \boldsymbol{\varepsilon}, \tag{2}$$

where $\mathbf{y}$ denotes measurements of the p-dimensional random variable $\mathbf{Y}$, ν is a p-dimensional vector of intercepts, Λ is a $p \times q$ dimensional matrix of regression coefficients or factor loadings on the q factors underlying $\mathbf{Y}$ $(p > q)$, η is the q-dimensional vector of factor scores, and, finally, $\boldsymbol{\varepsilon}$ is the p-dimensional residual score of the regression of $\mathbf{Y}$ on η.

In what follows, we assume that, in all subpopulations, $\mathbf{Y}$ is continuous, that the regression of $\mathbf{Y}$ on η is linear, and that the residuals have zero mean and are neither mutually correlated nor correlated with the factors. Note that the residual consists of the sum of two components, random error and a component which is specific to the observed variable $\mathbf{Y}$. Meredith's definition of SFI states that $\mathbf{Y}$ is strictly factorially invariant with respect to some selection variable V, if intercepts, ν, slopes, Λ, and residual variances, $\boldsymbol{\Theta}$, of the regression of $\mathbf{Y}$ on η are invariant across subpopulations derived by selection on V. Given SFI, means and covariances of $\mathbf{Y}$ in group i, where $i = 1, 2, \ldots, s$, can therefore be represented as follows:

$$\mu_i = \nu + \Lambda\alpha_i, \tag{3}$$

$$\Sigma_i = \Lambda\Psi_i\Lambda^{\mathrm{T}} + \boldsymbol{\Theta}. \tag{4}$$

Here, μ_i and α_i are vectors containing the means of $\mathbf{Y}$ and η in group i, respectively, and Σ_i, Ψ_i and $\boldsymbol{\Theta}$ represent the covariance matrices of $\mathbf{Y}$, η and $\boldsymbol{\varepsilon}$ in group i, respectively. Note that due to the restrictions of SFI, intercepts, ν, slopes, Λ, and residual variances, $\boldsymbol{\Theta}$, have no group index. Sörbom (1974) has shown how the mean and the covariance model can be estimated simultaneously using standard software for structural equation modelling. We assume that the matrix Λ contains sufficient fixed parameters to ensure identification of the factor model for as far as it concerns the covariance structure of the test scores. In addition, to ensure identification of the mean structure, it is necessary to arbitrarily choose one group, say the first, to be the reference group. The factor mean in the reference group, α_1, is fixed to zero. Hence, the estimated factor means in the remaining groups, $i = 2, \ldots, s$, can be interpreted in terms of factor mean differences with respect to the reference group.

SFI with respect to V almost certainly implies WMI with respect to V, because, given SFI, group differences in means and (co)variances of observed variables are almost certainly due solely to group differences in factor means and (co)variances (Meredith, 1993). The term 'almost certainly' is used in the previous sentences because of the indeterminacy concerning the residual variance mentioned above: SFI would certainly imply WMI if both specific and random error variance were equal across groups. In the factor model, only the sum of random and specific error is considered. Equality of the residual variance across subpopulations does not guarantee equality of both specific and random error variance across groups, only invariance of the sum of these two components. For instance, group differences in the specific component may indicate measurement bias because parameters of the factor model other than means and (co)variances of the factors depend on the selection variable V. However, in order to have equal residual variances, possible subpopulation differences in specific variance have to be accompanied by differences in random error variance and the two differences have to cancel each other out. Since this is very unlikely to be the case, Meredith

asserts that SFI with respect to V 'almost certainly' implies weak measurement invariance with respect to V.

In what follows, this result is extended. In groups selected on V, not only V but also variables that are correlated with V and/or with η can be excluded as possible sources of bias.

4. Weak measurement invariance with respect to unmeasured variables

We disregard special cases of selection on V which do not give rise to mean differences across groups (e.g. random selection). Rather, we focus on selection which results in differences in the expected value of V across subpopulations. Also, we assume that V and the latent variable underlying $\mathbf{Y}$ are stochastically dependent so that selection on V entails subpopulation differences in η and, consequently, in μ.

If groups are selected on V and SFI holds with respect to these groups, then we can take it as almost certain that V does not introduce bias. Suppose there is, in addition to the selection variable V and the latent variable, which is assumed to underlie the observed variable $\mathbf{Y}$, a potentially biasing variable W. At least some of the observed indicators have a non-zero regression coefficient on W. Furthermore, W is not measured and therefore not included in the model. The objective of the present paper is to examine the consequences of not including W in the model for groups selected on V if W introduces uniform bias in indicators of η (i.e. W has a main effect on one or more indicators), and if W introduces non-uniform bias (i.e. an interaction of W and η that can be modelled with a product term, see below). We argue that in both cases, SFI is not tenable in groups derived by selection on V if either W and η or W and V, or all three, are correlated. Consequently, if SFI is tenable, W does not introduce uniform or non-uniform bias under these conditions. To avoid possible misunderstandings we would like to emphasize that our argument pertains only to groups selected on V, and only to those biasing variables that are correlated with η and/or V. Importantly, we do not claim that SFI necessarily holds in groups selected on W. Let V again be sex, and W attitude towards mathematics, and assume these two variables to be correlated. If SFI holds in a maths achievement test across boys and girls, we argue that sex is not a biasing factor and that attitude to mathematics does not introduce bias in the two gender groups. It does not mean that if we create groups based on attitude to maths, SFI will hold for the mathematical achievement test across the attitude groups.

Following Mellenbergh (1982), we represent uniform bias by an additive main effect of the biasing variable W. For reasons of simplicity, we assume a biasing influence of W on one of the K indicators of η, say Y_k. Omitting group and subject subscripts, the score on Y_k equals

$$y_k = \nu_k + \lambda_{k\eta}\eta + \lambda_{kw}W + \varepsilon_k, \qquad k = 1, 2, \ldots, K, \tag{5}$$

where $\lambda_{k\eta}$ and λ_{kw} are the factor loadings of Y_k on the underlying factor and the unmeasured variable W, respectively. In case of non-uniform bias, Mellenbergh adds a product term, in our notation ηW, because the relation between η and Y_k depends on the level of W (see also Kenny & Judd, 1984). Importantly, the main effect of W remains in the model such that

$$y_k = \nu_k + \lambda_{k\eta}\eta + \lambda_{kw}W + \lambda_{k\eta w}\eta W + \varepsilon_k. \tag{6}$$

We will show that if the main effect of W or both main and interaction effect are

omitted from the model then SFI across subpopulations derived by selection on V is not tenable. It follows that tenability of SFI with respect to V (almost certainly) indicates absence of uniform and/or non-uniform bias by W in groups selected on V under the dependencies described above.

Conditional on V, the expected value of Y_k in the case of uniform and non-uniform bias is

$$E(Y_k) = \nu_k + \lambda_{k\eta} E(\eta) + \lambda_{kw} E(W), \tag{7}$$

$$E(Y_k) = \nu_k + \lambda_{k\eta} E(\eta) + \lambda_{kw} E(W) + \lambda_{k\eta w} E(\eta W), \tag{8}$$

respectively. In (7) and (8), $E(W)$ and $E(\eta W)$ indicate the expected values of W and the product variable, ηW, respectively. The variance of Y_k, $\mathrm{Var}(Y_k)$, under uniform and non-uniform bias, is

$$\mathrm{Var}(Y_k) = \lambda_{k\eta}^2 \mathrm{Var}(\eta) + \lambda_{kw}^2 \mathrm{Var}(W) + 2\lambda_{k\eta}\lambda_{kw} \mathrm{Cov}(W, \eta) + \mathrm{Var}(\varepsilon_k), \tag{9}$$

and

$$\begin{aligned}\mathrm{Var}(Y_k) = {} & \lambda_{k\eta}^2 \mathrm{Var}(\eta) + \lambda_{kw}^2 \mathrm{Var}(W) + \lambda_{k\eta w}^2 \mathrm{Var}(\eta W) + 2\lambda_{k\eta}\lambda_{kw} \mathrm{Cov}(\eta W, \eta) \\ & + 2\lambda_{k\eta}\lambda_{k\eta w} \mathrm{Cov}(W, \eta) + 2\lambda_{kw}\lambda_{k\eta w} \mathrm{Cov}(W, \eta W) + \mathrm{Var}(\varepsilon_k),\end{aligned} \tag{10}$$

respectively. The covariance between Y_k and some other, non-biased indicator of the underlying factor, say Y_l, equals

$$\mathrm{Cov}(Y_k, Y_l) = \lambda_{k\eta}\lambda_{l\eta} \mathrm{Var}(\eta) + \lambda_{l\eta}\lambda_{kw} \mathrm{Cov}(\eta, W) \tag{11}$$

in the case of uniform bias, and

$$\mathrm{Cov}(Y_k, Y_l) = \lambda_{k\eta}\lambda_{l\eta} \mathrm{Var}(\eta) + \lambda_{l\eta}\lambda_{kw} \mathrm{Cov}(\eta, W) + \lambda_{l\eta}\lambda_{k\eta w} \mathrm{Cov}(\eta, \eta W), \tag{12}$$

in the case of non-uniform bias. Note that, since Y_l is unbiased, the factor loading of Y_l on W, λ_{lw}, is zero, and all terms involving λ_{lw} are omitted in (11) and (12).

The distinct situations in which the unmeasured variable W is correlated with the selection variable (case 1, see Fig. 1), with the underlying factor (case 2, see Fig. 2), or with both (case 3, see Fig. 3) are considered separately. In addition, we briefly discuss the consequences of not including W in the model if W is stochastically independent of both η and V (case 4).

4.1. Case 1: W is correlated with V but stochastically independent of η

If W is correlated with V then the two variables are dependent and the expectation of W conditional on V is a function of V. Since subpopulations derived by selection on V differ with respect to V, there will be at least some subpopulations which differ also with respect to W (see Fig. 1). For instance, let V be race and W family income; then selection on race will induce mean differences in family income across at least some of the different ethnic groups. Stated otherwise, indirect selection on W due to direct selection on V introduces group differences in the conditional expectance of W given V, $E(W \mid V = v)$.

4.1.1. Uniform bias

Consider the expected value of Y_k if W is not included in the model. Subpopulation differences in $E(W)$ will be manifest in terms of differences in the intercept of Y_k as $\nu_k^* = \nu_k + E(W)$. As a result, the mean structure model with SFI restrictions as shown in

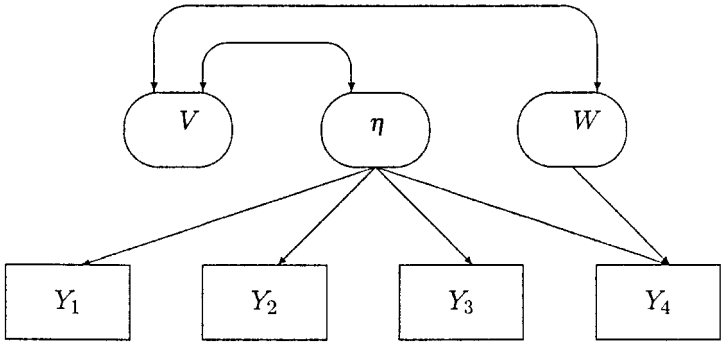

Figure 1. Path model representing case 1 in the present population. The unmeasured variable *W* and the selection variable *V* are stochastically dependent, and *W* and the factor underlying observed variables **Y** are stochastically independent.

(3) does not hold, because ν_k^* is not subpopulation invariant. If W is not included in the model, (3) under- or overestimates the mean difference in Y_k depending on the sign of the mean difference of W across subpopulations.

An exception occurs if the intercept of the regression of Y_k on η differs across groups but the sum of the group-specific intercept and $E_i(W)$ is equal across groups, that is, $\nu_i + E_i(W) = \nu$. However, this exception presupposes a second source of bias causing the intercepts to differ across groups and a compensation of the two biasing effects. We regard this exception as being of the same order as the indeterminacy in the residual which led Meredith (1993) to conclude that SFI 'almost certainly' implies weak measurement invariance.

Note that, generally, SFI will not hold even if all indicators of η are influenced by W. However, SFI will hold if the vector of loadings on η and the vector of loadings on W are collinear. In that case η and W cannot be distinguished and the estimate of $E(\eta)$ would not correctly represent the group difference in the factor. On a conceptual level, however, it seems difficult to imagine a biasing variable with linear relations to observed variables which are collinear to the relations of the factor one is interested in.

4.1.2. Non-uniform bias

The expected value of Y_k in the case of non-uniform bias is given in (8) and, as in the case of uniform bias, the mean structure model with SFI restrictions does not hold. As can be deduced from (8), the only exception here occurs if $\lambda_{kw} = \lambda_{kw}E(W) + \lambda_{k\eta w}E(\eta W)$.

Hence, if Y_k has a non-zero regression coefficient on an unmeasured potentially biasing variable W, and W is correlated with the selection variable V but stochastically independent of η, then SFI with respect to V is not tenable when omitting W from modelling.

The argument can be turned around. If SFI holds in groups selected on V, then the conditions of case 1 cannot hold. More specifically, if SFI holds in groups selected on V, then the conditional expectation of a potentially biasing variable W given V has to be invariant across all groups selected on V. Only if $E(W)$ is invariant under selection on V can the intercept of Y_k in (3) be replaced by $\nu_k^* = \nu_k + \lambda_{kw}E(W)$ without leading to a rejection of SFI in groups selected on V (given the additional assumption of invariance of λ_{kw} across groups). Furthermore, as can be deduced from (9) and (10), the

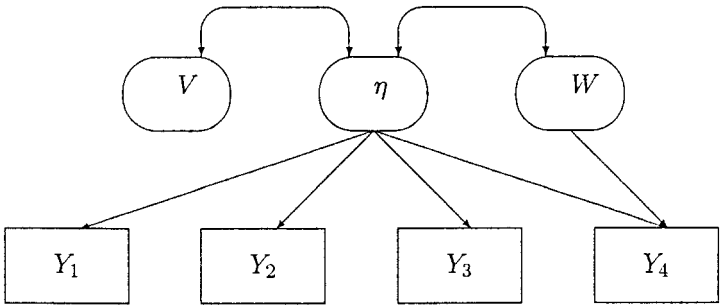

Figure 2. Path model representing case 2 in the parent population. The unmeasured variable *W* and the selection variable *V* are stochastically independent, and *W* and the factor underlying observed variables **Y** are stochastically dependent.

conditional variance of W given V has to be invariant for all realizations of V if SFI holds in groups selected on V. If the conditional variance is constant across groups, then the residual variance in (4) can be replaced by the sum of the residual variance and the conditional variance of W given V multiplied by the square of the respective factor loading, and SFI holds. In sum, tenability of SFI with respect to V almost certainly implies absence of uniform and/or non-uniform bias introduced by W in case 1.

4.2. Case 2: W is correlated with η but stochastically independent of V

The second part of our statement concerns the case where the unmeasured potentially biasing variable W and the selection variable V are stochastically independent variables, and there is prior knowledge that W and factor scores η are correlated (see Fig. 2). This differs from Case 1 in that the expectation of W conditional on V equals the unconditional expectation of W, meaning that there are no mean or variance differences in W across groups after selection on V. For instance, if sex and age are independent in the parent sample, selection on sex will not induce differences in age after selection.

4.2.1. Uniform bias

Assuming without loss of generality that the mean of W is zero in the parent population, (7) is consistent with (3): the mean difference in Y_k between groups is solely due to differences in η. But consider the (co)variances of Y_k with unbiased indicators of η if W is omitted. If SFI is tenable, according to (4) the covariance between Y_k and Y_l equals $\lambda_{k\eta}\lambda_{l\eta}\,\mathrm{Var}(\eta)$, which is inconsistent with (11).[3] Equation (11) can be rewritten as follows:

$$\mathrm{Cov}(Y_k, Y_l) = \lambda_{k\eta}\lambda_{l\eta}\,\mathrm{Var}(\eta) + \lambda_{l\eta}\lambda_{kw}\,\mathrm{Cov}(\eta, W) \tag{13}$$

$$= [\lambda_{k\eta} + \lambda_{kw}\,\mathrm{Cov}(\eta, W)/\mathrm{Var}(\eta)]\lambda_{l\eta}\,\mathrm{Var}(\eta) \tag{14}$$

$$= \lambda^{*}_{k\eta}\lambda_{l\eta}\,\mathrm{Var}(\eta). \tag{15}$$

[3] *If W and η are not jointly normally distributed, it is possible to construct cases in which the covariance between η and W is zero, but these covariances will differ from zero for at least some selection functions. If W and η are jointly normally distributed, the covariance always differs from zero.*

When applying (4) to groups selected on V, omission of W results in over- or underestimation of $\lambda_{k\eta}$, depending on the sign of the covariance between W and η and the signs of the loadings. Importantly, the factor loading $\lambda_{k\eta}$ in the covariance model differs from the corresponding parameter in the mean model, because W contributes to the covariances between Y_k and the unbiased Y-variables but not to the mean differences in Y_k. In this situation, (3) and (4) do not hold simultaneously, so SFI with respect to V is not tenable.

4.2.2. Non-uniform bias

The consequences of non-uniform bias introduced by W are as follows. If the effect of the product term ηW on the mean of Y_k varies across subpopulations, the result is straightforward: (3) does not hold (see (8)). If the effect is equal across the subpopulations under consideration, then (12) is rewritten as

$$\mathrm{Cov}(Y_k, Y_l) = \lambda_{k\eta}\lambda_{l\eta}\mathrm{Var}(\eta) + \lambda_{l\eta}\lambda_{kw}\mathrm{Cov}(\eta, W) + \lambda_{l\eta}\lambda_{k\eta w}\mathrm{Cov}(\eta, \eta w) \quad (16)$$

$$= \{\lambda_{k\eta} + [\lambda_{kw}\mathrm{Cov}(\eta, W) + \lambda_{k\eta w}\mathrm{Cov}(\eta, \eta w)]/\mathrm{Var}(\eta)\}\lambda_{l\eta}\mathrm{Var}(\eta) \quad (17)$$

$$= \lambda^{**}_{k\eta}\lambda_{l\eta}\mathrm{Var}(\eta) \quad (18)$$

(cf. (13)–(15)). Again, the factor loading in the covariance model differs from the corresponding loading in the mean model, and (3) and (4) will not hold simultaneously.

The conclusion in case 2 is the same as in case 1: uniform and/or non-uniform bias of Y_k with respect to W under the conditions of case 2 means that SFI with respect to V is not tenable. Consequently, SFI with respect to V is tenable only if the conditions of case 2 do not hold, that is, the biasing variable W is not correlated with η.

4.3. Case 3: W is correlated with η and V

Given the dependence relations of both case 1 and case 2, SFI with respect to V does not hold when omitting W due to specific discrepancies between the SFI model without W (i.e. (3) and (4)), and the corresponding models including W. Dependence of W and V resulted in over- or underestimation of the mean difference in Y_k whereas dependence of W and η resulted in an incorrect estimation of the factor loading of Y_k on η. The question might arise under which conditions these effects cancel out if W is correlated with both the selection variable and the latent variable η. If the discrepancies of cases 1 and 2 cancel out, SFI would hold and, consequently, bias would remain undetected.

4.3.1. Uniform bias

The discrepancy in case 1 between the model with and without W can be derived by substracting (3) from (7) and equals $\lambda_{kw}E(W)$. The discrepancy in case 2 equals the difference between (4) and (13), that is, $\lambda_{kw}\mathrm{Cov}(\eta, W)/\mathrm{Var}(\eta)$. Hence, if W is correlated with both V and η, SFI with respect to V holds if and only if

$$\lambda_{kw}E(W) = -\lambda_{kw}\mathrm{Cov}(\eta, W)/\mathrm{Var}(\eta) \quad (19)$$

$$E(W) = -\mathrm{Cov}(\eta, W)/\mathrm{Var}(\eta). \quad (20)$$

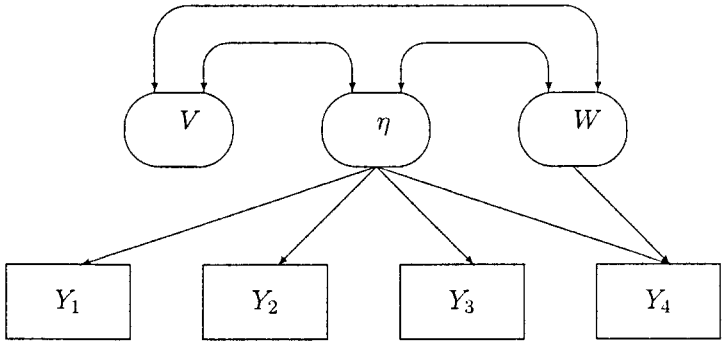

Figure 3. Path model representing case 3 in the parent population. The unmeasured variable *W*, the selection variable *V*, and the factor scores η are mutually dependent.

4.3.2. Non-uniform bias

The condition for non-uniform bias can be derived in a similar way:

$$\lambda_{kw}E(W) + \lambda_{k\eta w}E(\eta W) = -[\lambda_{kw}\,\mathrm{Cov}(\eta, W) + \lambda_{k\eta w}\,\mathrm{Cov}(\eta, \eta W)]/\mathrm{Var}(\eta), \qquad (21)$$

where the left-hand side is the discrepancy in case 1, that is, the difference between (3) and (8). The right-hand side gives the discrepancy in case 2, and equals the difference between (4) and (17).

Only if these rather special conditions are met do (3) and (4) hold simultaneously. If the above equalities do not hold and Y_k is uniformly and/or non-uniformly biased by W, SFI with respect to V is not tenable if W is omitted from modelling.

4.4. Case 4: W is stochastically independent of both η and V

If the biasing variable W and V are stochastically independent, selection on V will entail group differences neither in the mean nor in the variance of W. Hence, the influence of W on Y_k is invariant in subpopulations selected on V. The variance in Y_k due to W is absorbed by the residual variance of the regression of Y_k on the factor, but this residual variance remains invariant across groups. The covariance of W and η is zero in groups selected on V and does not disturb the covariance of Y_k and unbiased indicators of η. It follows that SFI with respect to V holds. Restricting the factor model according to SFI across groups selected on V is not a means to detect an influence of W if both W and η and W and V are stochastically independent. Suppose, for instance, that sex and ethnicity, and sex and an IQ factor are stochastically independent. Suppose further that selection is on ethnicity, and sex has an effect on one of the IQ test items. SFI with respect to ethnicity does not allow any conclusions with respect to the effect of sex if sex is stochastically independent of both V and η. Note however, that in the ethnic groups, the influence of sex is invariant. Now consider the following, somewhat more complicated example suggested by a reviewer.[4] In this example, we have a variable U in addition to W, V, and η (see Fig. 4).

U introduces bias and W is regressed on U, but U is stochastically independent of V and η. W is correlated with V. An example would be where V is ethnicity, W is family income, η is an IQ factor, and U is sex. Ethnicity is correlated with income but sex and ethnicity are independent, and, for the sake of this example, sex is again

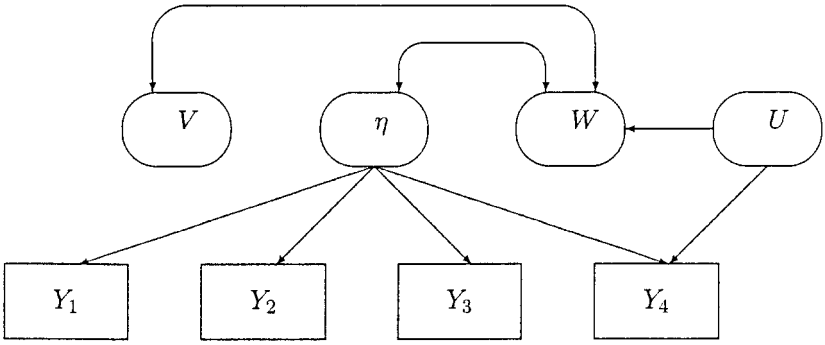

Figure 4. Path model representing the example provided by a reviewer. Note that the bias introduced by U might effect either Y_4 or the specific error of Y_4. The important feature is that the bias introducing variable is U, which is stochastically independent of V and η.

independent of the IQ factor. Income is regressed on sex. Let sex introduce bias in indicators of the IQ factor. Hence, if U is not observed, the biasing influence of U is conveyed via W. Again, SFI across ethnic groups is not a means to detect the influence of sex. The reason is that the variable U (sex) is stochastically independent of V and η (ethnicity and IQ, respectively). Our argument pertains only to those potentially biasing variables that are correlated with the selection variable and/or the underlying factor. In other words, if there is a set of potentially biasing variables, we state that SFI across groups selected on V allows conclusions to be drawn with respect to members of the subset of potentially biasing variables that are correlated with V and/or η but not with respect to those potentially biasing variables that do not exhibit these dependencies.

The results concerning tenability of SFI as presented in this section provide a theoretical basis for using the concept of SFI in practice. Tenability or rejection of SFI when fitting (3) and (4) simultaneously to data will depend on the size of sample estimates of correlations, factor loadings, etc. Therefore, an illustration of the three cases (e.g. W and V correlated, W and η correlated, and a combination of the two) based on simulated data is given in the next section. The illustration is restricted to uniform bias.

5. Illustration of cases 1–3 based on simulated data

5.1. Procedure

Case 1 represents the first part of the statement, which asserts that if an unmeasured biasing variable is correlated with the selection variable, then SFI does not hold. Case 2 concerns the second part where the biasing variable is correlated with the latent variable underlying the manifest biased variable. Case 3, finally, consists of a combination of case 1 and case 2. The general procedure used to illustrate cases 1–3 is as follows. The factor of interest, the unmeasured biasing variable, and the selection variable are unidimensional (see Figs 1–3). Sample data representing the parent population are

[4] *We would like to thank the reviewer for this contrived example which helps to clarify the scope of our argument.*

obtained by generating factor scores η, scores on W, and scores on V for n subjects. Regarding η, W, and V as factors, we use the common factor model (2) to compute manifest scores on $\mathbf{Y}$. The parent sample is divided into two subsamples by selection on V. Covariance matrices and mean vectors for the two subpopulation samples are computed and a two-group model incorporating the restrictions of SFI is fitted using normal theory maximum likelihood estimation (using LISREL 8.2). Tenability of SFI is evaluated by means of measures of goodness of fit and modification indices.[5]

The following characteristics are common to all three cases. The single underlying factor has four indicators, Y_1 to Y_4. The selection variable V and Y_1 to Y_3 are locally independent given the factor score η (see Figs 1–3). One of the four elements of $\mathbf{Y}$, Y_4, has a non-zero loading on W, meaning that Y_4 is uniformly biased with respect to W. The factor score η and variable W are both normally distributed with mean 0 and variance 1.5. The selection variable V is perfectly related to one dichotomous indicator taking values 1 and 0, which is a way of modelling an observed dichotomous selection variable such as sex. The residuals of the regression of $\mathbf{Y}$ on η, W, and V are normally distributed with $\varepsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{\Theta})$. The residual variance, $\mathbf{\Theta}$, depends on the reliability of $\mathbf{Y}$, which is varied (see below). The factor loadings of $\mathbf{Y}$ on η are [1 .55 .7 .65] throughout and the correlation between η and V equals .5. The sample sizes of the subsamples are $N_1 = N_2 = 1000$.

Besides the reliability of the elements of $\mathbf{Y}$, in all three cases the size of the loading of Y_4 on W is varied as well as the size of the correlation between η and W, and between V and W. As a measure of the reliability of $\mathbf{Y}$ we use the ratio of the true variance attributable to the common factors and the total variance of an indicator which is given in the LISREL output as the squared multiple correlation. The residual variances of the V-variables are chosen such that when fitting the true model (i.e. the model with which the data were generated), the mean reliability is either low (i.e. approximately .4) or appropriate (i.e. approximately .7). The size of the loading of Y_4 on W is either .3 or .5. As for the correlations, recall that in cases 1 and 3, the correlation between V and W is non-zero, whereas in case 2, V and W are independent and have therefore a zero correlation. In case 1, η and W are independent and have a zero correlation whereas in cases 2 and 3 the correlation is non-zero (see Figs 1–3, respectively). The non-zero correlations are varied and take the value .3 or .5. The data were generated using S-PLUS (4.5).

The data are analysed as follows. The correct model in the subpopulations is a two-factor model because the selection variable V is dichotomous in the parent population and, consequently, degenerate in subpopulations which are selected based on values of V. Fitting the two-factor model provides correct estimates of the subpopulation parameters. The relevant parameter estimates resulting from fitting the true model are given for each data set in the corresponding table to allow comparison. The focus of the illustration is on whether or not SFI is tenable if W is omitted from the comparison of the subpopulation data. This would happen in practice if W is not measured and, consequently, not included in the model. Hence, a single-factor model is fitted on the means and covariances of the four elements of $\mathbf{Y}$ simultaneously in the two subpopulations using (3) and (4), that is, the model with SFI restrictions. These restrictions entail that regression intercepts, factor loadings, and residual

[5] *Modification indices are measures of misspecification. A modification index indicates the decrease in χ^2 if a constrained parameter is freed and the model is re-estimated (Jöreskog & Sörbom, 1993). If the parameter equals the value as specified in the model, the corresponding modification index is χ^2-distributed with one degree of freedom.*

variances are equated across groups. Furthermore, the restrictions imply that the factor loadings are the same in the mean model and the covariance model. The only violation of SFI is due to omitting the influence of W on Y_4. More specifically, in case 1, V and W are dependent. As described in a previous section, selection on V implies indirect selection on W and the resulting difference in means of W across subpopulations leads in turn to mean differences of the residual of the regression of Y_4 on η. However, fitting (3) assumes that the mean differences in the residual are zero, which should result in evidence of misfit due to over- or underestimation of the mean difference in the biased variable Y_4. In case 2, the estimate of the factor loading of the biased variable on η will be incorrect, because the contribution of W to the covariances of Y_4 and the other Y-variables is partially conveyed through η. The over- or underestimated factor loading in turn will result in a failure to reproduce the mean difference in Y_4. In case 3, finally, underestimation and overestimation of the mean difference in Y_4 may tend to cancel each other out.

5.2. Results

5.2.1. Case 1

The mean difference in Y_4 in the simulated data equals the sum of mean differences in the factor and mean differences in W, each multiplied with the corresponding factor loading. The mean difference in Y_4 is underestimated by the mean model (3), because it does not take into account the contribution of W to the means of Y_4 when W is omitted from the model. This misfit can only partially be compensated by an overestimation of the factor loading of Y_4 on η, because η and W are stochastically independent and, consequently, W does not contribute to the covariances among the Y-variables. This limits the overestimation of the factor loading of Y_4 on η. The estimate of the factor loading appeared to be close to the true value with which the data were generated. Modification indices and measures of goodness of fit in Table 1 show that SFI was not tenable.

Even if the Y-variables are measured with low reliability and the loading of Y_4 on W as well as the correlation between V and W equals .3 in the parent sample (i.e. first row of Table 1), especially the modification index of the group difference in mean of Y_4, μ_δ demonstrates that SFI did not hold. In order to investigate whether the mean model is a source of misfit one can follow a two-step procedure (Mandys, Dolan, & Molenaar, 1994). In both steps, means and covariances are fitted simultaneously. In the first step, the mean model is unrestricted (i.e. $\mu_i = \nu_i$ in both groups). In the second step, the mean model is restricted to implement the full set of restrictions implied by SFI (3). The decrease in fit can be tested with a likelihood ratio (LR) test. In case 1 with reliability $\approx$.4 the difference in χ^2 is 15.07 and the difference in degrees of freedom equals 3. Hence, the LR test is significant with $p < .01$.

5.2.2. Case 2

Here, W contributes to the covariances of Y_4 with the other, unbiased Y-variables, which results in an overestimation of the factor loading of Y_4 on η when W is not included in the estimated model (see Table 2).

The overestimated factor loading in turn leads to an overestimation of the mean differences in Y_4. As in case 1, the data set with $\lambda_W = \rho_{VW} = .3$ and reliability $\approx$.4 is also analysed in a two-step procedure. The difference in χ^2 between the unrestricted and the

Table 1. Case 1 results

Parameter values parent population	Parameter estimates		Modification indices		Measures of model fit	
	$\lambda_{4\eta}$	α	$\lambda_{4\eta}$	$\mu_{\delta 4}$	χ^2 (p-value)	RMSEA
Reliability of observed Y-variables $\approx$.4						
$\lambda_{4W} = \rho_{VW} = .3$	.69 (.64)	.90 (.87)	3.58	15.17	29.77 (.01)	0.033
$\lambda_{4W} = \rho_{VW} = .5$	.74 (.64)	.94 (.88)	33.63	111.24	121.44 (.00)	0.090
Reliability of observed Y-variables $\approx$.7						
$\lambda_{4W} = \rho_{VW} = .3$	.67 (.65)	.93 (.87)	11.62	60.19	75.90 (.00)	0.066
$\lambda_{4W} = \rho_{VW} = .5$	.67 (.65)	.93 (.87)	73.95	313.46	338.20 (.00)	0.150

Note. Correct subpopulation parameter estimates derived by fitting a two-factor model in the two groups are given in parentheses.

Table 2. Case 2 results

Parameter values parent population	Parameter estimates		Modification indices		Measures of model fit	
	$\lambda_{4\eta}$	α	$\lambda_{4\eta}$	$\mu_{\delta 4}$	χ^2 (p-value)	RMSEA
Reliability of observed Y-variables $\approx$.4						
$\lambda_{4W} = \rho_{VW} = .3$	.74 (.64)	.84 (.86)	3.95	12.98	28.66 (.00)	0.032
$\lambda_{4W} = \rho_{VW} = .5$	.93 (.64)	.81 (.86)	15.77	58.39	78.59 (.00)	0.066
Reliability of observed Y-variables $\approx$.7						
$\lambda_{4W} = \rho_{VW} = .3$	.75 (.65)	.84 (.86)	9.21	28.50	46.57 (.00)	0.047
$\lambda_{4W} = \rho_{VW} = .5$	.96 (.65)	.83 (.86)	33.74	123.65	149.37 (.00)	0.095

Note. Correct subpopulation parameter estimates derived by fitting a two-factor model in the two groups are given in parentheses.

restricted mean structure equals 13.65, the difference in degrees of freedom equals 3, and the LR test is significant with $p < .01$. As in case 1, this result indicates a significant decrease in model fit when the mean model is restricted according to SFI. Case 2 shows that assessing tenability of SFI is a useful tool not only in case V and W are correlated, but also in case V and W are independent variables and W and η are dependent.

5.2.3. Case 3

In this case the effects of omitting W may tend to cancel each other out. As shown in Table 3, the results are less clear than in cases 1 and 2. Low reliability of the Y-variables, in combination with a small biasing effect of W, tends to obscure the violation of SFI.

Reanalysis of the data set with $\lambda_{4W} = \rho_{VW} = \rho_{W\eta} = .3$ and reliability $\approx$.4 using the two-step procedure does not show a significant decrease in model fit: the resulting χ^2 difference is 2.66.

It is noteworthy that the modification index regarding the mean model, $\mu_{\delta 4}$, is apparently more sensitive to violations of SFI than is χ^2. Even if the reliability of the

Table 3. Case 3 results

Parameter values parent population	Parameter estimates		Modification indices		Measures of model fit	
	$\lambda_{4\eta}$	α	$\lambda_{4\eta}$	$\mu_{\delta 4}$	χ^2 (p-value)	RMSEA
Reliability of observed Y-variables $\approx$.4						
$\lambda_{4W} = \rho_{VW} = \rho_{W\eta} = .3$	.76 (.65)	.88 (.86)	.50	7.00	17.33 (.24)	0.015
$\lambda_{4W} = \rho_{VW} = \rho_{W\eta} = .5$	.90 (.66)	.90 (.86)	5.02	19.90	34.95 (.00)	0.039
Reliability of observed Y-variables $\approx$.7						
$\lambda_{4W} = \rho_{VW} = \rho_{W\eta} = .3$	.75 (.65)	.88 (.86)	1.89	14.07	30.28 (.00)	0.034
$\lambda_{4W} = \rho_{VW} = \rho_{W\eta} = .5$	.89 (.66)	.90 (.86)	14.41	74.02	90.07 (.00)	0.074

Note. Correct subpopulation parameter estimates derived by fitting a two-factor model in the two groups are given in parentheses.

Y-variables is low, the modification index of $\mu_{\delta 4}$ in our example indicates that SFI is violated.

The results of the illustration can be summarized as follows. The dependence of V and W results in violation of SFI when W is not included in the model, because the dependence results in mean differences in Y_4 in the simulated data which are larger than those reproduced by the mean model. The mean model underestimates the mean differences in the data. Dependence of η and W, on the other hand, results in an overestimation of the true factor loading of Y_4 on η, which leads in turn to an overestimation of the mean difference in Y_4. If the dependence of V and W occurs simultaneously with independence of η and W and vice versa as in case 1 and case 2, respectively, SFI is rejected. In both cases, rejection is due to misfit of the mean model, and the results of cases 1 and 2 look surprisingly similar. In an empirical situation, cases 1 and 2 can be distinguished based on prior knowledge concerning dependence of V and W, and of η and W. Cases 1 and 2 are interesting mainly as limiting cases since independence of variables is the exception rather than the rule in the social sciences. When the two dependencies occur in combination as in case 3, the effects of omitting W from model fitting tend to partially cancel each other out. Nevertheless, the modification indices seem to be sensitive to violations of SFI. It is advisable to analyse the data in the two-step procedure suggested by Mandys *et al.* (1994) in order to detect violations of SFI which are due to discrepancies between the covariance and the mean model.

6. Conclusion

Establishing a tenability of SFI of a psychometric test with respect to a selection variable V has important consequences. In groups selected on V, one can exclude not only V as a biasing variable, but also all variables W that are correlated with V and/or with the latent variable underlying the test items or subscales. We have shown that restricting a factor model according to SFI is useful for detecting both uniform and non-uniform bias of W on a test item. Clearly, SFI can only be investigated through simultaneous modelling of means and (co)variances of the observed variable test scores (Meredith, 1993). Simultaneous modelling of means and (co)variances can be carried out with standard software for structural equation modelling such as LISREL. As stated by

Meredith (1993), it is important to realize that weak measurement invariance and strict factorial invariance are idealization; however, it should be evident that these are enormously useful concepts.

One issue that deserves attention concerns the origin of mean differences in the factor underlying **Y** when SFI is tenable with respect to V. Suppose that W is correlated with V and η (e.g. case 3). Tenability of SFI with respect to V implies that W has no influence on **Y** in addition to the influence that is conveyed through the factor (i.e. the loadings of **Y** on W are all zero). However, it is possible that η can be regressed on W. This would have no consequence for SFI if W is omitted from modelling. If there is such a regression relation, it is also possible that the group difference in the factor means is partially due to differences in W. For instance, let V represent sex, η verbal reasoning and W interest in reading. Suppose that sex and verbal reasoning are stochastically dependent and interest in reading boosts verbal reasoning, that is, the regression coefficient of η on W is positive. If SFI is tenable with respect to sex, then the verbal test is almost certainly not biased with respect to interest in reading, however, mean differences in the verbal reasoning factor may be at least partially due to differences in interest in reading. Tenability of SFI implies that mean differences across groups in observed scores are due to mean differences in the factors underlying the observed scores, but this says nothing about the scores of the factor mean differences.

It might be argued that tenability of SFI is rather unlikely in practice. However, Dolan (2000), using generally accepted measures of goodness of fit, showed that SFI was tenable in a comparison of representative samples of African and Caucasian Americans ($N_b = 306$, $N_w = 1868$) on the Wechsler Adult Intelligence Scale (Jensen & Reynolds, 1982). Moreover, the illustration provided in the present paper showed that even if SFI is rejected, investigating the composite hypothesis of SFI provides useful information about possible sources of misfit in the mean model and/or discrepancies between the mean and the covariance model.

Acknowledgements

The research of Gitta Lubke was supported through a subcontract to grant no. 5 R01 HD30995-07 by NICHD. The research of Conor Dolan was made possible by a fellowship of the Royal Netherlands Academy of the Arts and Sciences.

References

Bloxom, B. (1972). Alternative approaches to factorial invariance. *Psychometrika*, *37*, 425–440.

Dolan, C. V. (2000). Investigating Spearman's hypothesis by means of multi-group confirmatory factor analysis. *Multivariate Behavioral Research*, *35*, 21–50.

Ellis, J. L. (1993). Subpopulation invariance of patterns in covariate matrices. *British Journal of Mathematical and Statistical Psychology*, *46*, 231–254.

Jensen, A. R., & Reynolds, C. R. (1982). Race, social class and ability patterns on the WISC-R. *Personality and Individual Differences*, *3*, 423–438.

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika*, *36*, 409–426.

Jöreskog, K. G., & Sörbom, D. (1993). LISREL 8: User's guide. Chicago: Scientific Software International.

Kenny, D. A., & Judd, C. M. (1984). Estimating the nonlinear and interactive effects of latent variables. *Psychological Bulletin*, *96*, 201–210.

Mandys, F., Dolan, C. V., & Molenaar, P. C. (1994). Two aspects of the simplex model: Goodness of fit to linear growth curve structures and the analysis of mean trends. *Journal of Educational and Behavioral Statistics*, *19*, 201–215.

Mellenbergh, G. J. (1982). Contingency table models for assessing item bias. *International Journal of Educational Statistics*, *7*, 105–118.

Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research*, *13*, 127–143.

Meredith, W. (1964). Notes on factorial invariance. *Psychometrika*, *29*, 177–185.

Meredith, W. (1993). Measurement invariance, factor analysis, and factorial invariance. *Psychometrika*, *58*, 525–543.

Millsap, R. E. (1995). Measurement invariance, predictive invariance, and the duality paradox. *Multivariate Behavioral Research*, *30*, 577–605.

Millsap, R. E. (1997). Invariance in measurement and prediction: Their relationship in the single-factor case. *Psychological Methods*, *2*, 248–260.

Millsap, R. E. (1998). Group differences in regression intercepts: Implications for factorial invariance. *Multivariate Behavioral Research*, *33*, 403–424.

Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement*, *17*, 297–334.

Oort, F. J. (1992). Using restricted factor analysis to detect item bias. *Methodika*, *6*, 150–166.

Shealy, R., & Stout, W. (1993). An item response theory model for test bias. In P. Holland & H. Wainer (Eds), *Differential item functioning* (pp. 197–239). Hillsdale, NJ: Erlbaum.

Sörbom, D. (1974). A general method for studying differences in factor means and factor structure between groups. *British Journal of Mathematical and Statistical Psychology*, *27*, 229–239.

Received 31 January 2000; revised version received 27 May 2002